DREAM AND SLEEP ANALYSIS OF AN INDIVIDUAL

PROJECT DISSERTATION

SNEHA MUKHERJEE
ROLL-STI202015
REGI. NO-1348 of 2020-21
2yr. M.Sc Statistics

(Department of Statistics, Aliah University, Kolkata)

Under the Guidance of Sir. Hare Krishna Maity

CONTENTS

1. Summary

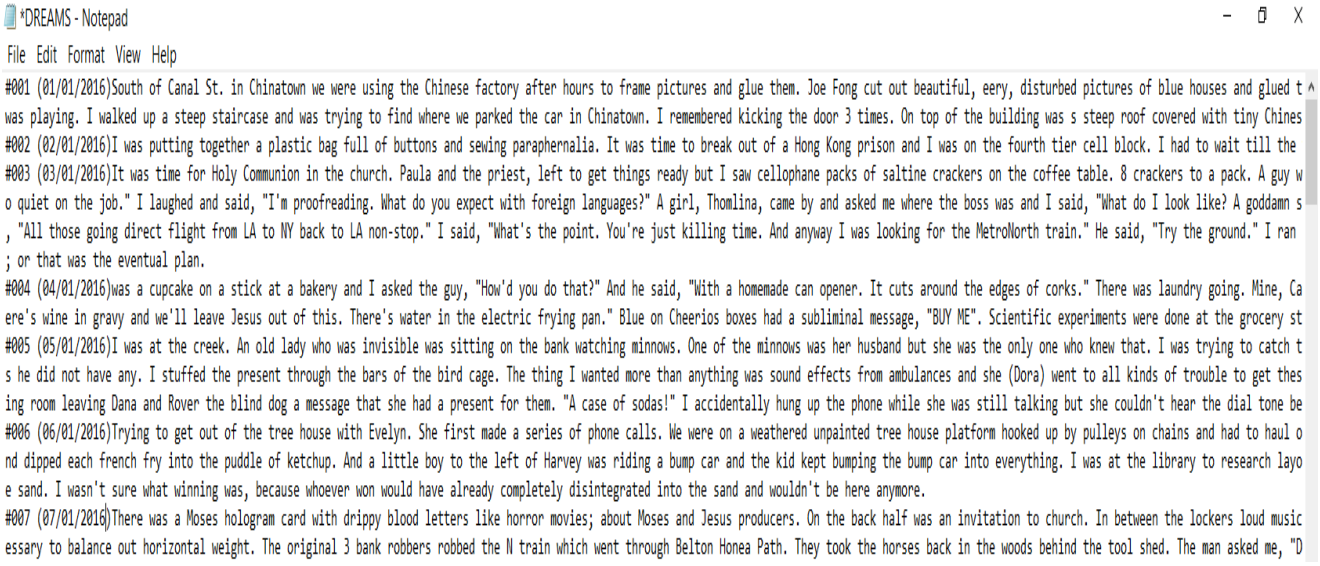
The aim of this analysis is to test if I can predict **the number of dreams which one person will recall** and **how happy those dreams will be and how his sleep quality is** using the datasets collected in a self-tracking app by him, I have used this sleep data and the dreams dataset, which he wrote in his phone's notepad, from the internet. .

2. Introduction-

In dream analysis , a dream is the beginning position for hypothesizing single features of the dreamer like personality quality or recent reviews. Most of the investigators (Strauch & Meier, Domhoff, Schredl) use dream content analysis as a process of analyzing dreams because the explanatory process has methodological issues. Dream content analysis is one of the important processes used in emotional dream study .This procedure has the benefit that it assures the general criterion of science such as replication by other research teams, measurement of reliability and authenticity, and minimizing bias of experiment.

3.. Summary of dataset-

I have a dream list in text format like this.



And I have a sleep dataset of an individual man which I have found from internet .

4. Dream Analysis -

- **Aim**-

I have to show how dream recall varies by weekdays and what emotions his dreams are made up of.

- **Dataset-**

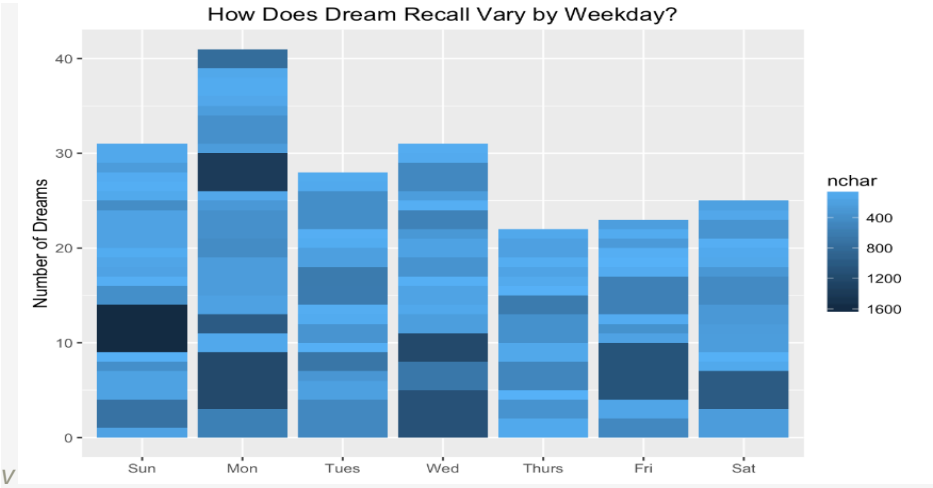
to show this I have taken the dataset into a csv file from that text file.

	dreams	Date
1	South of Canal St. in Chinatown	1.1.2016
2	I was putting together a plastic	2.1.2016
3	It was time for Holy Communio	3.1.2016
4	was a cupcake on a stick at a bal	4.1.2016
5	I was at the creek. An old lady w	5.1.2016
6	this. I said, "Alan, this is an awfu	6.1.2016
7	Trying to get out of the tree hou	7.1.2016

4.1. Part I: Studying the data based on plot and trying to find dream recall case

To examine the two options for dependent variables - proxies for dream recall, I have used the bar plot.

- **Visualization-**



- **R code-**

```
dreamcount <- dreamsplus %>% group_by(Date) %>%
  summarize(count = n(),
            nchar = sum(nchar))
ggplot(dreamcount, aes(x= wday(Date, label = T, abbr = T), y = count, fill = nchar)) +
  geom_bar(stat = "identity") +
  xlab("") +
  ylab("Number of Dreams") +
  ggtitle("How Does Dream Recall Vary by Weekday?") +
  scale_fill_continuous(trans = 'reverse')
```

- **Interpretation–**

The person recalled the most dreams on Monday nights. Thursday, Friday, and Saturday nights have the lowest dream counts. This is especially telling since my dataset goes from January 1st,2016 (a Friday night) to last night (a Saturday night), so we could look for a small raise to those weekdays. The number of characters he writes down on a given morning ranges from 0 to ~1600. Long dreams appear pretty spread out throughout the week.

4.1.1. Prediction of How many dreams he will remember-

Now I will show if he can predict the number of dreams he will recall or not.

- **Statistical modeling-**

For this prediction, I will use multiple linear regression.

Multiple linear regression–

Multiple linear regression (ML or multiple regression, is a statistical method which uses different descriptive variables to forecast the output of a response variable. It is an expansion of linear (OLS) regression which applies only one explanatory variable

Formula and Calculation of Multiple Linear Regression—

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots .. \beta_p x_{ip} + \epsilon$$

where,

y_i = Dependent variable

x_i = Explanatory variables

β_0=y-intercept

β_p= slope coefficients for every explanatory variable

ϵ= the model's error term (known as the residuals)

Assumptions of The multiple regression model –

There is a linear relationship among the independent and dependent variables

- The independent variables are not too extremely correlated with each other
- y_i observations are chosen independently and radically from the population
- Residuals should be normally distributed with a mean of 0 and variance σ

The coefficient of determination (R-squared) is a statistical term which measures how much of the variation in result can be described by the variation in the independent variables. R^2 always raisses as more predictors are included in the MLR model, although the predictors may not be connected to the result variable.

- **Step-**

By this multiple linear regression model, I will forecast how many dreams he will recall using the sleep duration and what day of week it is.

- **Result-**

By doing this,I can get this result

```
## Call:
## lm(formula = Count ~ Duration + DayofWeek, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4564 -0.9355 -0.1484  0.6998  5.0302
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.45186    0.81326   0.556    0.579
## Duration     0.13860    0.10403   1.332    0.185
## DayofWeek   -0.07654    0.05146  -1.487    0.139
##
## Residual standard error: 1.109 on 129 degrees of freedom
## Multiple R-squared:  0.03813,    Adjusted R-squared:  0.02321
## F-statistic: 2.557 on 2 and 129 DF,  p-value: 0.08149
```

- **Interpretation-**

From this analysis, we can say neither day of week nor duration are significant in predicting how many dreams he will recall. Together they can only say 2.3% of the variation in dream memory.

- **R code-**

```
x = lm(Count ~ Duration + DayofWeek, data)
summary(x) #Not a significant model
```

4.1.2. Prediction on whether he has a good dream or not-

From this dataset, I will also try to forecast his dream nature if it is good or not.

- **Procedure-**

To find this I will use a decision tree

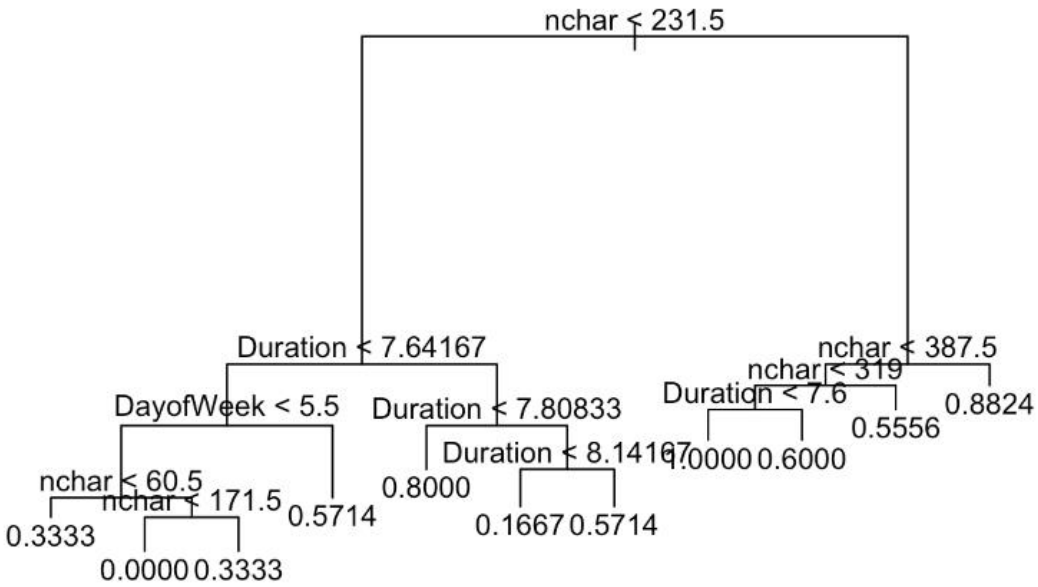
Decision tree-

A decision tree is a very prominent probability tree which helps us to create a decision about some kind of procedure. It is a map of the feasible outputs of a sequence of related options. It shows a person to consider plausible measures against one another dependent on their charges, advantages and probabilities.

Decision trees create regression or classification models in the form of a tree structure. It divides a dataset into smaller sub sets whereas at the same time an connected decision tree is gradually expanded. The final output is a tree with leaf nodes and decision nodes.

- ***Visualization –***

Now have shown the classification decision tree



- **R code-**

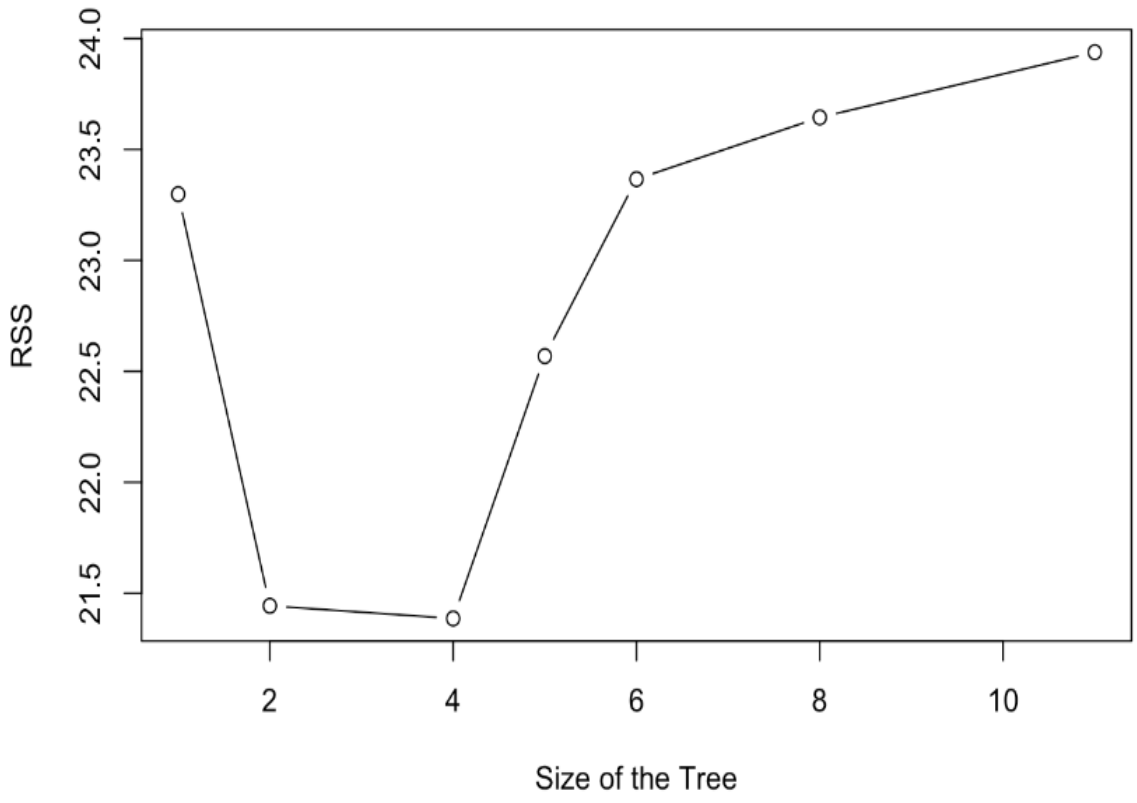
```
library(tree)

tree_model = tree(Joy ~ DayofWeek + Duration + nchar, corrdata, method = "class")
plot(tree_model)
text(tree_model, pretty=0)
```

Pruning in decision tree-

Pruning is a data shortening process in machine learning and explore algorithms which shrinks the size of decision trees by eliminating segments of the tree which are reliable and unnecessary to organize cases.

- **Visualization-**



- **R code-**

```
#Pruning the tree
cv.tree <- cv.tree(tree_model)
plot(cv.tree$size,
     cv.tree$dev,
     type = "b",
     xlab = "Size of the Tree",
     ylab = "RSS")
pruned_tree = prune.tree(tree_model, best=4)
```

- **Result–**

From these ,4 nods create the lowest RSS (measure of deviation/ measure of error).

RSS-

RSS(residual sum of squares) is a statistical method shown to calculate the variance in a data which is not described by a regression model itself. Instead, it estimates the variance in the residuals, or error term. It is also known a known as the sum of squared residuals or the sum of squared estimate of errors or sum of squared residuals.

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

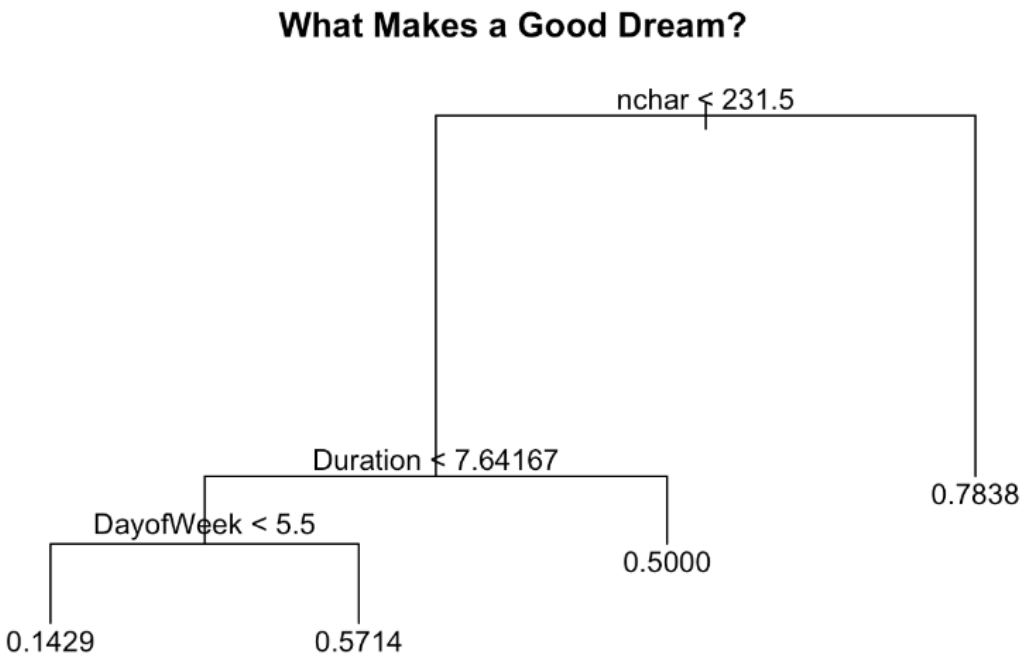
RSS = residual sum of squares

y_i = i^th value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

- **Visualization-**



- **R code–**

```
pruned_tree = prune.tree(tree_model, best=4)
plot(pruned_tree)
text(pruned_tree)
title("What Makes a Good Dream?")
```

- **Interpretation-**

Showing the tree model: When he writes more than ~230 characters, there's a 78% possibility he will show joy. If he wrote less than that, the likelihood that he has a good dream relies on how long he sleeps and on what day of the week it is. He will be unable to have a joyful dream if he doesn't get a lot of sleep and it's a weekday.

4.2. Part 2: finding the emotions of dreams

To show this, I will use the sentiment analysis of the dreams using 'sentiment total' in R programming language.

Sentiment analysis in R-

The packages which are used in sentiment analysis in are

- tm for text mining operations like deleting numbers, special characters, punctuations and stop words (Stop words in any language are the generally occurring words which have very small value for NLP(natural language processing) and should be cleaned out
- word cloud for creating the word cloud plot.
- syuzhet for sentiment scores and classification of emotion
- ggplot2 for plotting and showing graphs
- sentimentTotals food showing the emotions like anger, anticipation,iissggust, fear, joy,negative etc.

- **Expected result-**

After getting this file, I added the sentiment columns and got a summary of it .

```
## Observations: 201
## Variables: 14
## $ Date         (time) 2016-01-01, 2016-01-01, 2016-01-02, 2016-01-02, ...
## $ Dream        (chr) "At a dinner with some soccer and lax girls and t...
## $ Date_simple  (chr) "1/1", "1/1", "1/2", "1/2", "1/2", "1/3", "1/4", ...
## $ nchar        (int) 198, 272, 142, 80, 19, 175, 225, 174, 164, 78, 29...
## $ anger        (dbl) 0, 0, 1, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0...
## $ anticipation (dbl) 1, 0, 0, 0, 0, 0, 2, 1, 1, 0, 0, 0, 0, 3, 0, 1, 1...
## $ disgust      (dbl) 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0...
## $ fear         (dbl) 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0...
## $ joy          (dbl) 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 2, 0, 0, 1, 0, 0, 0...
## $ sadness      (dbl) 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0...
## $ surprise     (dbl) 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0...
## $ trust        (dbl) 1, 0, 0, 1, 0, 2, 0, 1, 1, 0, 1, 0, 0, 2, 0, 0, 1...
## $ negative     (dbl) 1, 1, 1, 0, 0, 0, 2, 0, 0, 0, 1, 0, 0, 2, 0, 0, 0...
## $ positive     (dbl) 3, 1, 0, 1, 0, 2, 1, 1, 2, 0, 2, 0, 0, 2, 2, 0, 1...
```

- **R code-**

```
library(syuzhet)
library(ggplot2)
sentiments <- get_nrc_sentiment(dreams$Dream)
dreamsplus <- cbind(dreams, sentiments)
glimpse(dreamsplus)
```

- ***Actual result–**

I got the sentiments which were contained in dream dataset.

```
##    count    sentiment
## 1    219     positive
## 2    135        trust
## 3    126     negative
## 4    121 anticipation
## 5    101          joy
## 6     77         fear
## 7     69     surprise
## 8     63      sadness
## 9     49        anger
## 10    41      disgust
```

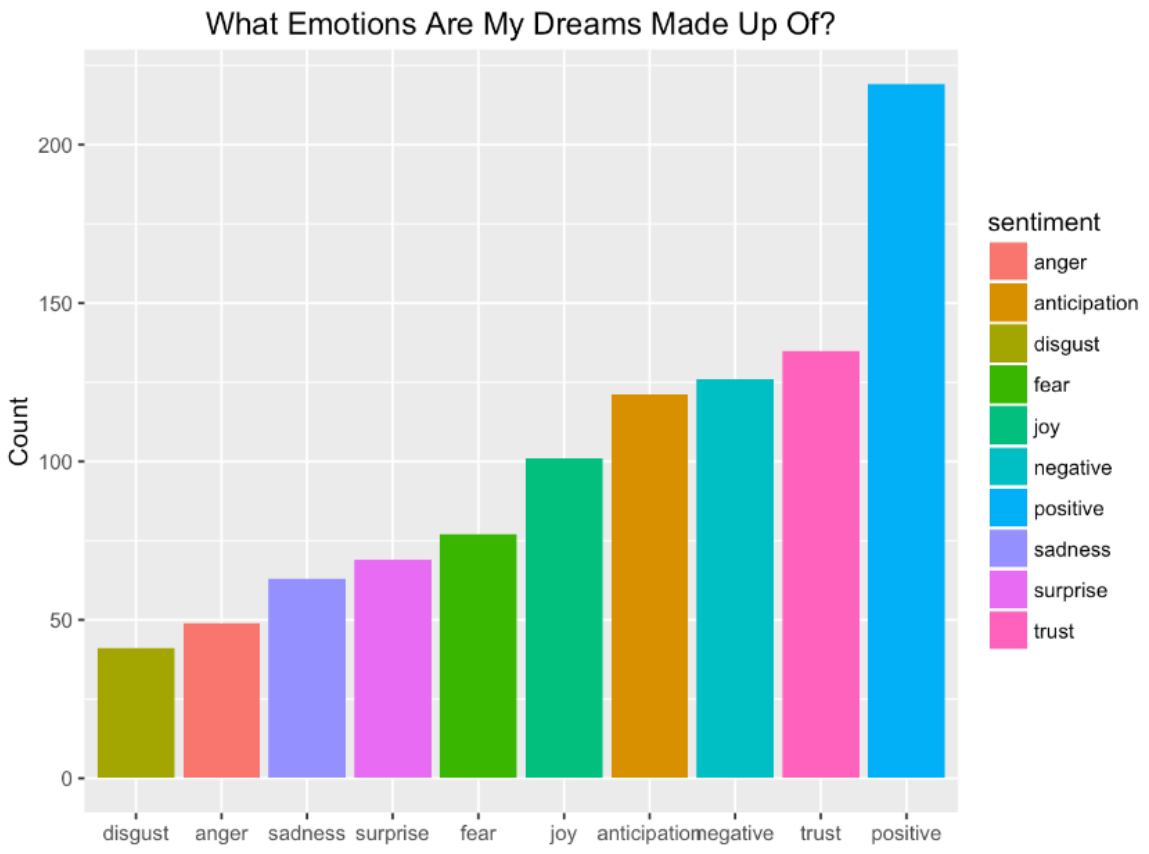
- **R code-**

```
sentimentTotals <- data.frame(count = colSums(sentiments))
sentimentTotals$sentiment = rownames(sentimentTotals)
arrange(sentimentTotals, -count)
```

- **Interpretation-**

'Positive' is by far the most common sentiment chosen by the Syuzhet package. After doing an easy spot check with the text itself, it appears to be generously assigned. I'll use 'Joy' as the proxy for a good dream.

- **Visualization**



- **R code-**

```
ggplot(sentimentTotals, aes(x = reorder(x = sentiment, count), y = count, fill = sentiment)) +
  geom_bar(stat = "identity") +
  theme(axis.title.x = element_blank()) +
  labs(y = "Count", title = "What Emotions Are My Dreams Made Up Of?")
```

5. Sleep analysis–

Now I have taken a sleep dataset to show how dream and sleep are interrelated.

- **Why this is needed-**

If his sleep is proper or not, how much time he sleeps at night and wake up in the morning— from this analysis, we can find all the details.

- ***Scope of analysis -**

To do this, I have cleaned the raw dataset by changing sleep quality and removing percentage sign etc. Next I have created a date column, because sometimes he sleeps after midnight, so this is the most proper date to match up with dream date.after that I will remove all the unused column (e.g. excluding nappss- the observations less than 2 hours)from the dataset.

- **Dataset** -

I have got this data from a sleep cycle app of a person's sleep record.

Start	End	Quality	Duration	heart r	Activit	Date	date simple
19.1.2016	20.1.2016	78.00	6.63	NA	14825	19.1.20	1/19
20.1.2016	21.1.2016	65.00	7.12	NA	7793	20.1.20	1/20
21.1.2016	22.1.2016	63.00	7.28	NA	10793	21.1.20	1/21
22.1.2016	22.1.2016	78.00	7.05	NA	15560	22.1.20	1/22
23.1.2016	23.1.2016	81.00	8.13	NA	4314	23.1.20	1/23
24.1.2016	25.1.2016	73.00	5.23	NA	4561	24.1.20	1/24

- **Result**-

From this dataset, I have done the needed changing and got the result as

```
## Observations: 146
## Variables: 8
## $ Start       (time) 2016-01-19 23:33:50, 2016-01-21 00:00:32, 2016-01...
## $ End         (time) 2016-01-20 06:12:08, 2016-01-21 07:08:09, 2016-01...
## $ Quality     (dbl) 78, 65, 63, 78, 81, 73, 82, 70, 70, 79, 62, 83, 78...
## $ Duration    (dbl) 6.633333, 7.116667, 7.283333, 7.050000, 8.133333, ...
## $ Heart.rate  (int) NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ Activity    (int) 14825, 7793, 10790, 15560, 4314, 4561, 10576, 8579...
## $ Date        (time) 2016-01-19 06:12:08, 2016-01-20 07:08:09, 2016-01...
## $ Date_simple (chr) "1/19", "1/20", "1/21", "1/22", "1/23", "1/24", "1...
```

- ***R code-**

```
sleep$Quality <- as.numeric(sub("%","",sleep$Quality))
#change dates and times from text to POSIXct
sleep$Start <- ymd_hms(sleep$Start)
sleep$End <- ymd_hms(sleep$End)
#Create simple date column
sleep$Date <- sleep$End - days(1)
sleep$Date_simple <- paste(month(sleep$Date), day(sleep$Date), sep = "/")
#Change sleep duration to decimal hours i.e. 7.5 for 7 hours 30 minutes
sleep$Time.in.Bed<- as.numeric(as.duration((sleep$Time.in.Bed)))/3600

#Rename columns for ease of use, remove unused columns
sleep <- rename(sleep, Duration =sleeptime, Activity = Steps)
sleep <- filter(sleep, Duration > 2
glimpse(sleep)
```

- **Analysis -**

Now I have created a heatmap to show the analysis of getting up in the morning and going to bed at night.

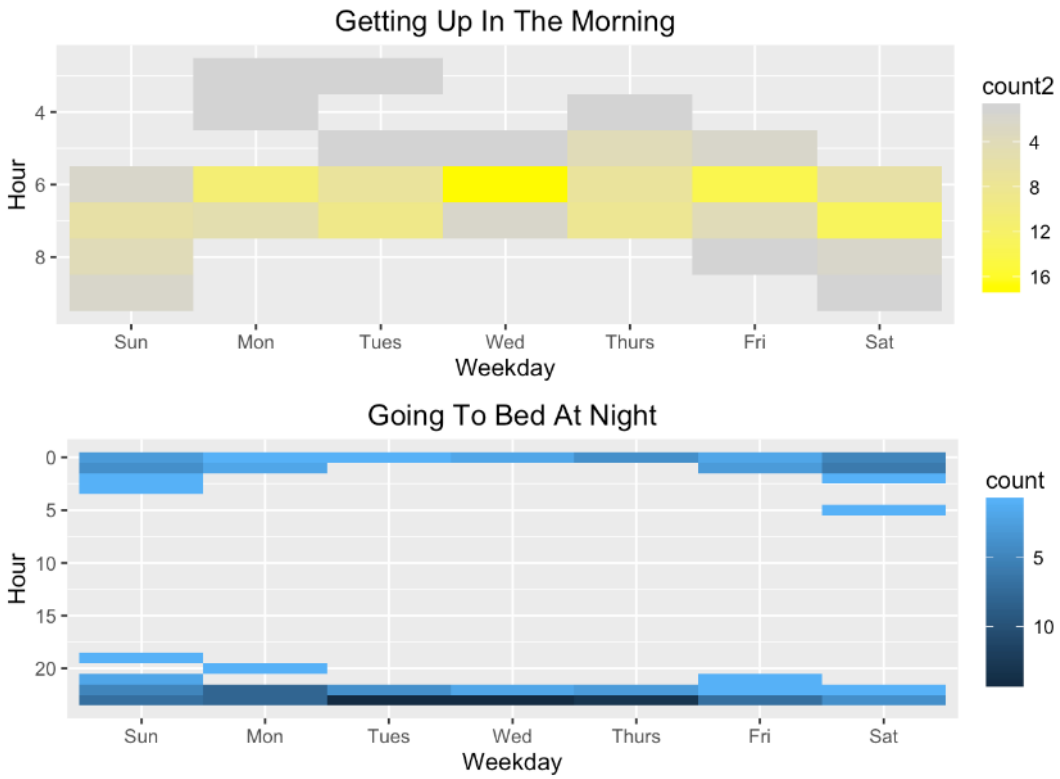
Heatmap-

A heatmap operates color to display modifications and magnitude of a third variable to a two-dimensional graph. It is used to visualize design and changes. whereas it can be used to show transformation over time, heatmap is a display of ssom rectangles. The x-axis is frequently some measure of time but it can be any variable with category. The y-axis is a variable which describes the categories in data. Every rectangle is the equal size. The rectangles are colored to explain the magnitude of the third variable. Though it was first used for temperatures, heatmaps are now used for many types of data. Heatmaps are important for big data sets. Heatmap with a time axis can show

patterns and changes over time. Heatmap rectangles can be labeled with values of the color variable, which is useful only in cases where there are very few categories on the y-axis.

From this heatmap I can say,the time he goes to sleep differs more than the time he wakes up, as provided by the different persons. So from this analysis I can say that he is a morning person and wake up before time even when he goes to sleep behind time, frequently followed by a nap in the afternoon.

- **Visualization-**



- **R code–**

```
gotosleep <- sleep %>% mutate(weekday = wday(sleep$Start, label = T), hour = hour(sleep$Start)) %>%
  group_by(weekday, hour) %>%
  summarize(count = n())

wakeup <- sleep %>% mutate(weekday2 = wday(sleep$End, label = T), hour2 = hour(sleep$End)) %>%
  group_by(weekday2, hour2) %>%
  summarize(count2 = n())
ggstart = ggplot(gotosleep, aes(x = weekday, y = hour, fill = count)) +
  geom_tile() +
  ggtitle("Going To Bed At Night") +
  xlab("Weekday") +
  ylab("Hour") +
  scale_fill_continuous(trans = 'reverse') +
  scale_y_reverse()


ggend = ggplot(wakeup, aes(x = weekday2, y = hour2, fill = count2)) +
  geom_tile() +
  ggtitle("Getting Up In The Morning") +
  xlab("Weekday") +
  ylab("Hour") +
  scale_fill_continuous(low = "yellow", high = "lightgrey", trans = 'reverse') +
  scale_y_reverse()

#put the plots on top of each other
library(gridExtra)
grid.arrange(ggend, ggstart)
```

6. Merging of two datasets-

After merging 2 dataset ,now I have merged from this dream and sleep dataset into one to see the days which he has seen dreams at night, those days if he can sleep properly or not.

- **Expected Result-**

I have to find the correlation between sleep quality and sleep duration and between the number of dreams he has recalled and the number of characters he wrote .

- **Actual result-**

I found this observation

```
## Observations: 132
## Variables: 10
## $ Date_simple (chr) "1/21", "1/28", "1/28", "2/3", "2/5", "2/9", "2/10...
## $ Date        (time) 2016-01-21 06:37:49, 2016-01-28 06:07:21, 2016-01...
## $ Quality     (dbl) 63, 70, 70, 90, 69, 74, 64, 75, 70, 70, 75, 78, 51...
## $ Duration    (dbl) 7.283333, 6.233333, 6.233333, 8.266667, 5.816667, ...
## $ Heart.rate  (dbl) NA, NA, NA, 61, 75, 55, 66, 51, NA, NA, 46, 84, 96...
## $ Count       (int) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ nchar       (int) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Joy         (dbl) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ DayofWeek   (dbl) 5, 5, 5, 4, 6, 3, 4, 5, 6, 6, 2, 3, 4, 3, 4, 7, 1,...
## $ DayofMonth  (int) 21, 28, 28, 3, 5, 9, 10, 11, 12, 12, 15, 16, 17, 2...
```

- **Steps**-

To do this I have used correlation matrix..

Correlation Matrix:

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all possible pairs of values in a table. Now for our data, we have checked whether our data, independent variables are correlated or not (multicollinearity present or not). For this we compute correlation matrix for our numerical independent variables.

- **R code –**

For merging dataset–

```
#No dreams - Assign a count, nchar, and joy of 0 for dates in the sleep dataset that do not have a match
##in the dream dataset, so that I can use one table for analysis
sleep$Count <- match(sleep$Date_simple, table = dreams$Date_simple, nomatch = 0)
sleep$nchar <- match(sleep$Date_simple, table = dreams$Date_simple, nomatch = 0)
sleep$Joy <- match(sleep$Date_simple, table = dreams$Date_simple, nomatch = 0)
#Create subset of night with 0 dreams
nodreams <- filter(sleep, Count == 0) %>%
  select(Date_simple, Date, Quality, Duration, Pressure, Count, nchar, Joy)

#ndreams - Summarize the number of dreams and total characters for dates that appear in both datasets
ndreams <- merge(dreamsplus, sleep,
                 by.x = "Date_simple",
                 by.y = "Date_simple") %>%
  filter(!is.na(Duration)) %>%
  group_by(Date_simple) %>%
  summarize(Date = mean(Date.y),
            Quality = mean(Quality),
            Duration = mean(Duration),
            Count = n(),
            nchar = sum(nchar.x),
            Joy = sum(joy))
#combine the subsets
data <- bind_rows(nodreams, ndreams)
#Add date-related variables and categorical Joy for final dataset
data$DayofWeek = wday(data$Date)
data$DayofMonth = day(data$Date)
data$Joy = ifelse(data$Joy > 0, 1 , 0)
glimpse(data)
```

To get the correlation plot-

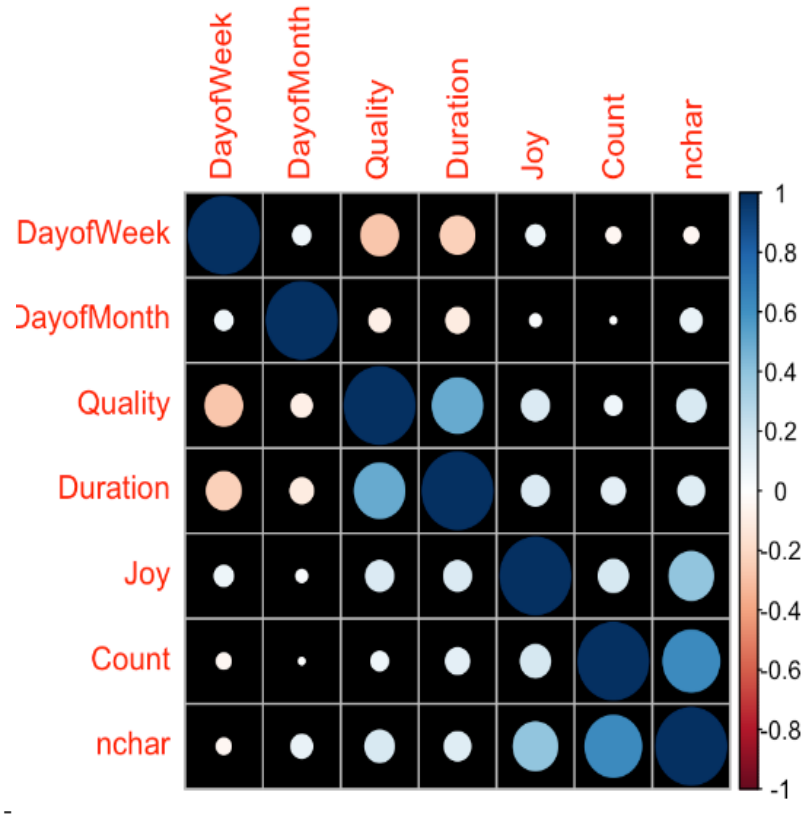
```
library(corrplot)

corrdata = data %>% select(-Date_simple, -Date) %>% filter(Count > 0)

correlations = cor(corrdata)
corrplot(correlations, order="hclust",
         bg = "black",
         title = "How Are All of These Variables Correlated?",
         mar = c(2,2,2,2))
```

- **Visualization**

How Are All of These Variables Correlated?



-

- **Interpretation-**

Showing the correlation plot: overlook the deep blue circles going diagonally-each variable has an ideal correlation with itself. The larger the dot, the stronger the correlation is. Reed is a negative correlation, and blue is positive correlation.

Here I am only seeing at nights where he recalled at least one dream, so that we don't get a deceptively big correlation between joy and dream recall. Sadly I also had to eliminate heart rate because he started gathering it fairly recently so there are many NAs.

There is a good correlation between sleep duration and sleep quality, and between the number of dreams he remembers and the number of characters he wrote. Sleep duration and quality both reduce over the course of the week. There is no correlation with the time of the month. Joy is most influenced by the number of characters he wrote - the more he remembers and write, the more likely it is that he'll recall a joyous dream. This is probably true for every sentiment.. not a tremendous imminent.

7. DISCUSSION :

I did not obtain many workable perceptions to remember more dreams or have better dreams, but I learned a lot about the analytics procedure. From this project, I could be able to find the emotions included in his dreams and can say how his sleep quality is and how dream and sleep are correlated among each other.

8. Future Work-

In future I will gather more datasets for different persons. Then I will find the comparison between their dreams . I will also find out if gender differences are necessary for dream interpretation or not.

9. ACKNOWLEDGEMENT

It has been a great opportunity to gain lots of experience in Logistic Regression followed by machine learning. We would like to express our special thanks to our project supervisor Sir Hare Krishna Maity who gave us the golden opportunity to do this wonderful project.

10. Reference-

- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale: Lawrence Erlbaum
- Hall, C. S., & Van de Castle, R. L. (1966). The content analysis of dreams. New York: Appleton-Century-Crofts
- Roussy, F., Brunette, M., Merrier, P., Gonthier, I., Grenier, J., Sirois-Berliss, M., Lortie-Lussier, M., & De Koninck, J. (2000). Daily events and dream content: unsuccessful matching attempts. Dreaming, 10, 77-83
- Zane, M. D. (1971). Significance of differing responses among psychoanalysts to the same dream. In J. H. Masserman (Ed.), Dream dynamics (pp. 174-177). New York: Grune & Stratton.Zane, M. D. (1971).
- Significance of differing responses among psychoanalysts to the same dream. In J. H. Masserman (Ed.), Dream dynamics (pp. 174-177). New York: Grune & Stratton

—----------xxxx xxxx—----------